Environmental agreements under asymmetric information*

Aurélie Slechten†

Abstract

In a two-country model, I investigate the role of a pre-negotiation phase as an information-sharing and certification device to restore the feasibility of an efficient environmental agreement when countries' abatement costs are private information and participation is voluntary. When uncertainty regarding abatement costs is high, the welfare gains of reducing information asymmetries and reaching the first-best agreement will be sufficiently large to design budget-balanced transfers that compensate both countries for the loss of the information rent they could obtain by staying privately informed. Both countries then accept to share and certify their abatement costs during the pre-negotiation phase.

Keywords: environmental agreements; asymmetric information; certification; information exchange

JEL Codes: Q54, D82

*This research has received funding from the FRS - FNRS. I am especially grateful to Estelle Cantillon, for her very helpful comments and constructive discussions. I am also indebted to Nicolas Sahuguet, Patrick Legros, Paola Conconi, Andreas Lange and David Martimort for their comments. I also thank the participants of the internal seminar at ECARES, the 11th Journées André-Louis Gérard Varet, the 21th Annual Conference of the EAERE, the 2014 Conference on Auctions, Competition, Regulation and Public Policy and the 18th meeting of the Association for Public Economic Theory. Finally, I thank two anonymous referees and the editor for very useful recommendations in improving the paper. All remaining errors are my own.

†Lancaster University Management School, a.slechten@lancaster.ac.uk

The classical explanation for the failure of international negotiations on environmental issues is the free-rider problem: countries have the possibility to opt out of the negotiations while still enjoying the benefits of the global agreement. If they have some private information, countries may also have an incentive to exaggerate their privately known costs of implementing the agreement (or understate their privately known benefits) in order to reduce the effort they have to supply and leave most of the burden on other countries. It is well-known from the mechanism design literature that inefficiencies arising from the free-rider problem are particularly relevant in contexts plagued by information asymmetry.[1] At the same time, international environmental agreements are very often preceded by discussion rounds, during which countries do not negotiate quantitative targets but exchange information regarding the costs and benefits of an agreement. This paper takes a mechanism design approach and studies the effect of asymmetric information about pollution abatement costs on the feasibility of an efficient environmental agreement when participation in this agreement is voluntary. It also investigates the role of a pre-negotiation phase as an information-sharing device to alleviate the inefficiencies generated by information asymmetry and voluntary participation.

Environmental agreements differ in scope and substance but most of them tend to be formulated through a process following a similar pattern (Wagner 2001). Countries first agree on an initial convention, i.e. an *umbrella convention*, that generally does not contain any emission reduction target or any monetary transfer scheme between countries. Rather, this convention sets up institutions entitled to negotiate all the subsequent environmental agreements with binding commitments. In 1979, following increasing concerns by policy-makers about the harmful effects of acid rains, 33 countries signed the Convention on Long-Range Transboundary Air-Pollution (LRTAP). This initial Convention served as a basis for eight follow-

[1]For example, Rob (1989) and Mailath & Postlewaite (1990) stress the role of participation constraints to generate inefficiency.

up protocols and a series of amendments, but it did not mention any numerical goal or abatement measure. Other examples include the United Nation Framework Convention on Climate Change (1992) or the Vienna Convention for the Protection of the Ozone Layer (1985), which acted as frameworks for subsequent agreements containing quantitative targets, i.e. the Kyoto and Montreal protocols respectively.

A key component of this pre-negotiation phase is the creation of a scientific body to investigate particular issues. It may also involve the assessment of existing legal regimes. For example, the LRTAP Convention put in place a structure to gather information on national emissions as well as national pollution and energy policies. An emission monitoring system was also set up under the auspices of the United Nations Economic Commission for Europe (UNECE) to verify information transmitted by the countries.[2] The contribution of this paper is twofold. First, I show how the presence of information asymmetry about pollution abatement costs may exacerbate the free-rider problem and result in the implementation of inefficient agreements. The second and main contribution is the introduction of a pre-negotiation stage during which countries have the opportunity to exchange verifiable information through an *international agency* (assisted by a scientific body) and reduce the information asymmetry at the negotiation stage. The objective is to examine in which contexts countries have an incentive to use this pre-negotiation phase as an information-sharing device and whether this restores the feasibility of an efficient agreement.

My analysis is carried out within the framework of the private provision of a

[2]This is also true for the two other examples mentioned above. By signing the UNFCCC, industrialized countries committed themselves to provide the Conference of the Parties with clear data about their greenhouse gas (GHG) emissions and about regional programs containing measures to mitigate climate change and with information related to implementation (which could give an indication of the political willingness to implement emission reductions). A subsidiary body of the UNFCCC was in charge of assessing this information. This reporting obligation, however, did not cover the developing countries. The Vienna Convention established the United Nations Environment Program (UNEP) as a secretariat and asked this body to *convene a workshop to develop a more common understanding of factors affecting the ozone layer, including the costs and effects of possible control measures* (Benedick 1998, p. 45).

public good under asymmetric information. The public good considered here is the reduction of some transboundary pollutant. Countries incur emission abatement costs and can only capture a share of the environmental benefits generated by their abatement efforts. Environmental negotiations are modeled as a two-stage process. In a first stage, countries sign an umbrella convention and set up an international agency. They agree on the role and the prerogatives of this agency, but as it is the case for most existing umbrella conventions, they don't agree on any emission targets or monetary transfers. In a second stage, countries negotiate an environmental agreement consisting of binding commitments to some emission abatement levels and monetary transfers. In the first stage, they only have common prior beliefs about their emission abatement costs and they have to decide on the information structure of the negotiation stage. Specifically, they have two options. First, they can investigate these costs individually and report the results of these investigations to the international agency, which will then design an agreement based on that information. In that case, emission reductions will be negotiated under asymmetric information. Second, as it was the case under the LRTAP convention, they could set up a framework for sharing information (through an international agency assisted by a scientific body, for example). To be useful in reducing asymmetric information, this second option requires that countries can rely on the results of the joint investigations about abatement costs and can base their negotiations about quantitative targets in the next stage on these results. In other words, the international agency should serve as a *certification* device, i.e. it should have the ability to monitor and verify the information transmitted by countries.[3]

I first consider a model without pre-negotiation stage, in which each country knows its own abatement cost, but not that of the other country. An agreement is feasible if all countries are willing to participate and if they all reveal their abatement

[3]The role of certification device played by the international agency is important. In a model without certification, a country could possibly manipulate strategically the information that is being shared.

cost truthfully. Due to a trade-off between ensuring participation in the agreement and truth-telling, a first-best agreement, i.e. an agreement that maximizes global welfare, is not feasible when the range of the distribution of abatement cost types is too large. A large support means that abatement costs are more uncertain and also that potential abatement costs are more heterogenous. With homogeneous abatement costs (more concentrated support), incentives to misreport its own type are less important, so it becomes easier to solve the tension between truth-telling and participation to implement the first-best agreement.

Second, I introduce a (pre-negotiation) certification stage, during which countries can decide to share information through an international agency, which will then certify countries' types. The effects of certification at the negotiation stage are twofold. On the one hand, the information asymmetry between countries is reduced. Intuitively, this effect will be stronger when the level of uncertainty about abatement costs is high because the proportion of types exerting their non-cooperative abatement effort in the second-best agreement increases. For this reason, the total welfare under the second-best agreement will be close to its non-cooperative level, which results in substantial welfare gains of reducing information asymmetries and reaching the first-best agreement. On the other hand, the country certifying its abatement cost type loses the possibility to misreport its type, and thereby may see its monetary transfer in the first-best agreement reduced. This effect will be weaker for higher levels of uncertainty because the share of types receiving an information rent in the second-best agreement (i.e. types that do not exert the non-cooperative abatement effort) is smaller.

Consequently, certification will restore the feasibility of the first-best agreement only if the level of uncertainty is high because the first effect dominates and both countries agree to gather information through the international agency. For those high levels of uncertainty, the welfare gains of reducing information asymmetries and reaching the first-best agreement will be sufficiently large to design first-best budget-

balanced transfers that compensate both countries for the loss of their expected information rent. For lower levels of uncertainty, a larger proportion of types is entitled to an information rent in the second-best mechanism and it is impossible to design first-best budget-balanced transfers for which both countries at the same time opt for certification. There is always at least one country free-riding on the other country's certification.

Some articles have developed specific applications of the mechanism design theory to environmental economics (e.g. Rob 1989; Baliga & Maskin 2003; Caparros *et al.* 2004; Helm & Wirl 2014; Konrad & Thum 2014 or Helm & Wirl 2016a). The model the closest to the one developed in this study is that of Martimort & Sand-Zantman (2016). They also highlight a trade-off between solving free riding due to asymmetric information and voluntary participation. They derive the conditions under which the first-best abatement levels are not implementable and analyze the characteristics of a second-best mechanism. Similar inefficiencies were pointed out in related setups under some conditions regarding the payoff functions and the distributions of types.[4] By contrast, I propose a channel to restore the feasibility of the first-best agreement, i.e. the introduction of a pre-negotiation stage with information exchange and certification.

In the contexts of transboundary pollution and incomplete information, some articles have analyzed how unilateral actions can be used as a signaling device to transmit countries' private information and either influence abatement decisions of other self-interested countries (Brandt 2004; Elofsson 2007), or affect the outcome of future environmental negotiations (Espinola-Arredondo & Munoz-Garcia 2012;

[4]In the context of public good economies, Laffont & Maskin (1979) show that no truthful and efficient mechanisms may exist if individual rationality constraints are taken into account and if, without transfers, some types are worse off under the efficient mechanism than under the outside option. This occurs when the heterogeneity in terms of payoffs is too large. In the private goods case, Myerson & Satterthwaite (1983) show the impossibility of attaining ex-post efficiency with an incentive-compatible and individually rational mechanism if the seller's and buyer's valuations are distributed over intervals that have a nonempty intersection. If these intervals intersect, it is more uncertain whether the valuation of the seller is higher than the valuation of the buyer.

Brandt & Nannerup 2013). In this paper, I focus on the role of information sharing and certification (at an ex-ante stage, abstracting from signaling incentives) in the formation of efficient environmental agreements. Kakeu & Johnson (2018) also study the incentives for information-sharing at an ex-ante stage. However, they concentrate on a different problem in which the unknown damage cost is the same for both countries but each of them receives a private signal about this unknown parameter and uses it to form non-cooperatively its environmental policy.

The remainder of the article is organized as follows. Section 1 lays out the main assumptions of the two-country model. Section 2 shows the effect of asymmetric information on the feasibility of the first-best agreement without certification. The results of this section (i.e. the existence of a trade-off between solving free riding due to asymmetric information and voluntary participation) are along the lines of the existing literature, e.g. Martimort & Sand-Zantman (2016). Section 3 is the main contribution: the introduction of a pre-negotiation stage with certification. Section 4 discusses some assumptions of the model. I conclude in section 5.

1 Setting of the model

There are two countries or two groups of countries ($i = 1, 2$) that exert some non-negative pollution abatement efforts a_i. Country i's payoff derived from abatement activities in both countries is given by:

$$\frac{1}{2}(a_1 + a_2) - \frac{a_i^2}{2\theta_i}$$

Global abatement benefit is simply the total quantity of abatement, i.e. $(a_1 + a_2)$. This global benefit is shared equally between countries: each country i receives a share of this global benefit equal to $(a_1 + a_2)/2$. Countries are heterogeneous in terms of their marginal cost of abatement. By exerting abatement effort a_i, country

i incurs a cost of $\frac{1}{2\theta_i}a_i^2$. For tractability, I adopt a quadratic form where θ_i can be interpreted as the characteristic of the technology of country i. A higher θ_i corresponds to a lower cost of abatement effort. In the rest of the paper, we will use the term *low-cost countries* to refer to countries with a high θ_i.

Environmental agreements between countries consist of abatement levels a_i^k and transfers t_i^k for each country: $y^k = (a_1^k, a_2^k, t_1^k, t_2^k)$. Country i may receive a transfer t_i^k for undertaking the requested abatement (where the superscript k is used to differentiate between the types of agreements analyzed in the model, e.g. first-best, second-best, etc.). Examples of international treaties allowing for the possibility of monetary or technology transfers are the Montreal Protocol or the Kyoto Protocol.[5]

As this will reveal useful for the analysis, I define two important benchmark cases: the non-cooperative equilibrium and the first-best agreement. When countries don't negotiate an international agreement, they choose their abatement levels to maximize their own payoff. As they act non-cooperatively, there is no transfer between countries.

Definition 1 *The non-cooperative abatement levels* (a_1^N, a_2^N) *are the abatement levels that maximize each country's payoff:*

$$a_i^N \in argmax_{a_i} \frac{1}{2}(a_1 + a_2) - \frac{a_i^2}{2\theta_i}$$

And $y^N = (a_1^N, a_2^N, t_1^N, t_2^N)$ *is the non-cooperative (Nash) equilibrium where* $a_i^N = \frac{1}{2}\theta_i$ *and* $t_i^N = 0$*, for* $i = 1, 2$*. The payoff of country* i *when both countries choose their non-cooperative abatement levels is:* $\frac{1}{8}\theta_i + \frac{1}{4}\theta_j$*, for* $i \neq j$*.*

Under the non-cooperative equilibrium, countries do not internalize the bene-

[5]Article 10 of the Montreal Protocol established a fund to facilitate technological cooperation about non-ozone depleting substances and technology transfers to assist developing countries. Article 11 of the Kyoto Protocol allows for the possibility of monetary transfers from developed to developing countries.

fits of their abatement activity on the other country and there is an under-provision of emission abatement. The second benchmark case is the first-best agreement in which global welfare is maximized and countries internalize the impact of their own abatement choice on the other country:

Definition 2 *The first-best abatement levels* (a_1^{FB}, a_2^{FB}) *are the abatement levels that maximize the global welfare:*

$$(a_1^{FB}, a_2^{FB}) \in argmax_{a_1,a_2} \ (a_1 + a_2) - \frac{a_1^2}{2\theta_1} - \frac{a_2^2}{2\theta_2}$$

And $y^{FB} = (a_1^{FB}, a_2^{FB}, t_1^{FB}, t_2^{FB})$ *is the first-best agreement where* $a_i^{FB} = \theta_i$ *and* $t_i^{FB} \in \mathbb{R}$ *are such that* $t_1^{FB} + t_2^{FB} = 0$*, for* $i = 1, 2$*. The payoff of country* i *when both countries choose their first-best abatement levels is:* $\frac{1}{2}\theta_j + t_i^{FB}$*, for* $i \neq j$*.*

In both benchmark cases, low-cost countries are those that abate the most. It is also interesting to note that the increase in abatement efforts required by the first-best agreement compared to the non-cooperative equilibrium, $a_i^{FB} - a_i^N = \frac{1}{2}\theta_i$, is more pronounced for low-cost countries (i.e. countries with a higher θ_i).

In our setting, benefits from abatement are linear, which substantially simplifies the analysis and is an assumption widely used in the literature.[6] This implies that a_i^N and a_i^{FB} are dominant strategies (i.e. the abatement chosen is the same whatever the behavior of the other country). This separability of efforts of course neglects incentives to free-ride on the environmental benefits of cooperation. However, this problem is less severe here as we are looking at agreements in which both countries participate.

[6]Finus *et al.* (2006) show that a linear specification can be justified for substantive reasons since discounted climate damages that are linear in emissions are a good approximation of the figures in the RICE model (Nordhaus & Yang 1996).

2 Introducing asymmetric information

I introduce information asymmetry by assuming that the cost parameter θ_i is privately observed by each country i.[7] As it is usually assumed in the mechanism design literature, the countries' types θ_i are independently drawn from the same uniform distribution, defined on the support interval $R_\theta = [\theta_M - \gamma, \theta_M + \gamma]$.[8,9] $\theta_M > 0$ is the average abatement cost type. The length of the support is given by 2γ, with $\gamma \in (0, \theta_M)$. In the rest of the paper, we will interpret γ as a parameter that measures the level of uncertainty about abatement costs (in the sense of mean-preserving spread). The cumulative and probability distribution functions are respectively denoted by $F(\theta_i) = \frac{\theta_i - (\theta_M - \gamma)}{2\gamma}$ and $f(\theta_i) = \frac{1}{2\gamma}$.

Mechanisms

In this setting with information asymmetry, agreements between countries are modeled using the concept of *mechanism*. By the revelation principle, there is no loss of generality in restricting our attention to direct and truthful revelation mechanisms (Myerson 1982). A direct revelation mechanism, $y^k = (a_1^k, a_2^k, t_1^k, t_2^k)$, is composed of a level of abatement for each country $a_i^k : R_\theta \times R_\theta \to \mathbb{R}$ that describes the abatement effort of each country as a function of countries' reported types, and a transfer

[7] Those costs should be understood in a broad sense as including not only technological costs but also the opportunity and political costs of achieving a given emissions target. Therefore, even though the abatement technologies might be the same across countries, the implementation costs may differ. Also, as pointed by Espinola-Arredondo & Munoz-Garcia (2012), international environmental agreements usually target overall emission levels, requiring the adoption of clean technologies by several industries in a country's economy, the precise dissemination of these technologies along all industries is difficult to observe by outsiders. This dissemination can, however, be more accurately assessed by local governments.

[8] All the results in section 2 hold under more general distributions of types (see the proofs in Appendices A1 and A2). The uniform distribution is made to obtain closed-form solutions in section 3.

[9] A number of results in the mechanism design literature show that when agents' types are correlated, private information is irrelevant(full rent extraction, see McAfee & Reny 1992). However, in the context of voluntary public good provision, Neeman (2004) has shown that full rent extraction results do not hold if uncertainty about an agent's type is in fact two dimensional. For example, if uncertainty about a country's type includes both uncertainty about the country's abatement cost and uncertainty about its beliefs and if countries' beliefs do not uniquely determine their abatement costs, then the extraction of the countries' entire informational rents is impossible.

$t_i^k : R_\theta \times R_\theta \to \mathbb{R}$ that describes each country's received transfer from undertaking the requested abatement effort as a function of countries' reported types.

I denote the utility of a country with type θ_i from the direct revelation mechanism $y^k(\hat{\theta}_1, \hat{\theta}_2)$, where $\hat{\theta}_i$ are the reported types for each country i, by:

$$V_i(y^k(\hat{\theta}_i, \hat{\theta}_j)|\theta_i) = \frac{1}{2}(a_i^k(\hat{\theta}_i, \hat{\theta}_j) + a_j^k(\hat{\theta}_i, \hat{\theta}_j)) - \frac{a_i^{k^2}(\hat{\theta}_i, \hat{\theta}_j)}{2\theta_i} + t_i^k(\hat{\theta}_i, \hat{\theta}_j)$$

To be implementable under asymmetric information and voluntary participation, a mechanism must satisfy three constraints. First, the mechanism must be incentive compatible. Second, the mechanism must ensure that both countries want to join the agreement (i.e. individual rationality). Finally, there is also a budget balance constraint. These constraints are detailed below.

Bayesian incentive compatibility. Bayesian incentive compatibility of the mechanism $y^k(.,.)$, requires that the expected utility of country i satisfies:

$$\theta_i = argmax_{\hat{\theta}_i \in R_\theta} E_{\theta_j}[V_i(y^k(\hat{\theta}_i, \theta_j)|\theta_i)] \tag{1}$$

where $E_{\theta_j}[.] = \int_{R_\theta}[.]f(\theta_j)d\theta_j$ denotes the expectation over the possible types of country j. In other words, truth-telling gives country i the highest possible expected utility, provided the other country j does. As, in the sequel, I focus on incentive compatible direct revelation mechanisms, I will simplify the notation of the expected utility of country i from the direct and truthful revelation mechanism $y^k(\theta_i, \theta_j)$ when this country i is of type $\theta_i \in R_\theta$:

$$U_i^k(\theta_i) \equiv E_{\theta_j}[V_i(y^k(\theta_i, \theta_j)|\theta_i)]$$

A country of type θ_i will have an incentive to exaggerate its cost of effort (e.g. by reporting a type $\hat{\theta}_i < \theta_i$) because it can abate at the same level as a higher-cost

country $\hat{\theta}_i$ but at a lower marginal cost. Therefore, low-cost countries have more incentives to misreport. To ensure truth-telling (condition (1)), the mechanism must reward the countries with the lowest abatement costs by an extra amount that is just equal to the gains from slightly overstating their abatement cost. In Lemma 1, I show that this extra amount is given by expression (2).

Lemma 1 *The direct revelation mechanism $y^k(.,.)$ satisfies Bayesian incentive compatibility if and only if $E_{\theta_j}[a_i^{k^2}(\theta_i,\theta_j)]$ is weakly increasing in θ_i and*

$$\dot{U}_i^k(\theta_i) = \frac{E_{\theta_j}[a_i^{k^2}(\theta_i,\theta_j)]}{2\theta_i^2} \tag{2}$$

Where $\dot{U}_i^k(.)$ is the derivative of the expected utility function of country i (from the direct and truthful revelation mechanism y^k) with respect to θ_i.

Proof See Online Appendix A1. □

From Lemma 1, it immediately follows that an incentive compatible mechanism must give a greater payoff to countries with lower abatement costs. Indeed, integrating equation (2) yields:

$$U_i^k(\theta_i) = U_i^k(\theta_M - \gamma) + \int_{\theta_M-\gamma}^{\theta_i} \frac{E_{\theta_j}[a_i^{k^2}(s,\theta_j)]}{2s^2} ds$$

where $U_i^k(\theta_M - \gamma)$ is the expected utility of country i when it has the highest abatement cost (i.e. the country that has no incentive to misreport its type) and the second term on the right-hand side is the additional payoff required by a country of type θ_i to ensure truth-telling.

Under the first-best agreement, $a_i^{FB} = \theta_i$ and $U_i^{FB}(\theta_i) = \frac{1}{2}\theta_M + E_{\theta_j}[t_i^{FB}(\theta_i,\theta_j)]$(see Definition 2). Using Lemma 1, we can then derive the first-best transfers that will satisfy Bayesian incentive compatibility:

$$E_{\theta_j}[t_i^{FB}(\theta_i,\theta_j)] = \frac{1}{2}[\theta_i - (\theta_M - \gamma)] + E_{\theta_j}[t_i^{FB}(\theta_M - \gamma,\theta_j)] \tag{3}$$

where $E_{\theta_j}[t_i^{FB}(\theta_M - \gamma, \theta_j)]$ is the first-best expected transfer when country i has a type $\theta_i = \theta_M - \gamma$ and the term $\frac{1}{2}[\theta_i - (\theta_M - \gamma)]$ can be interpreted as the *information rent* necessary to ensure truthful revelation by all types $\theta_i > \theta_M - \gamma$. This term is increasing in θ_i because low-cost countries contribute more to the total abatement under the first-best agreement (we have seen previously that $a_i^{FB} - a_i^N$ is increasing in θ_i). The coefficient $\frac{1}{2}$ in expression (3) is equal to the difference between the social and private marginal benefits of abatement and can be interpreted as a per-unit Pigouvian subsidy granted to a country for undertaking the first-best abatement effort.

Finally, we can see from equation (3), that the information rent is increasing in the level of uncertainty, γ. When γ is small, the support R_θ is concentrated around the mean, θ_M. Countries are relatively homogeneous and incentives problems are less severe. All countries have similar abatement levels under the first-best agreement and the incentives to misreport are lower. By contrast, when γ increases, there are more opportunities for low-cost countries to misreport their types. Avoiding such free-riding will require larger information rents.

Budget balance. I assume that no external source of funds is available and that there is no waste of resource.[10] Hence, the ex-ante budget-balance constraint is:

$$\int_{R_\theta} \int_{R_\theta} [t_1^k(\theta_1, \theta_2) + t_2^k(\theta_1, \theta_2)] f(\theta_1) f(\theta_2) d\theta_1 d\theta_2 = E[t_1^k(\theta_1, \theta_2) + t_2^k(\theta_1, \theta_2)] = 0 \quad (4)$$

where $E[.]$ denotes the expectation over the set of possible types for countries 1 and 2. Intuitively, the mechanism must be self-financed. Any transfer to a country must be covered by actual contributions. Note that there is no loss of generality in using the ex-ante budget balance constraint instead of the more natural ex-post budget balance constraint ($t_1^k(\theta_1, \theta_2) + t_2^k(\theta_1, \theta_2) = 0$) because, following Börgers & Norman

[10] I can relax this budget balance constraint, i.e. by assuming linkages with agreements in other areas, but the main results would remain unchanged.

(2009), if types are independent, for every ex-ante budget-balanced mechanism, there exists an ex-post budget-balanced mechanism such that the allocation rule is unchanged and the interim expected payments are unchanged for all agents.

Individual rationality. Finally, participation in an environmental agreement y^k is voluntary. The outside option is the non-cooperative equilibrium (see Definition 1). Since countries know their type when deciding to join a treaty, the expected utility of country i under this outside option is denoted by:

$$U_i^N(\theta_i) \equiv E_{\theta_j}[V_i(y^N(\theta_i, \theta_j)|\theta_i)] = \frac{1}{8}\theta_i + \frac{1}{4}\theta_M \tag{5}$$

The interim individual rationality constraint then requires that:

$$U_i^k(\theta_i) \geq U_i^N(\theta_i) \quad \text{for } i = 1, 2 \tag{6}$$

Implementation of the first-best agreement

Before turning to the analysis of the two-stage game, I will first examine whether the first-best abatement levels can be implemented under asymmetric information about abatement costs and voluntary participation in an agreement. This will help define when a pre-negotiation stage with a certification device can be used as a channel to restore the feasibility of an efficient agreement. Combining Bayesian incentive compatibility (2), budget balance (4) and interim individual rationality (6) yields the following proposition:[11]

Proposition 1 *The first-best agreement y^{FB} is implementable for all $(\theta_1, \theta_2) \in R_\theta \times R_\theta$ if and only if the support interval of the distribution of types, R_θ, satisfies:*

$$\gamma \leq \frac{\theta_M}{3} \tag{7}$$

[11] This is the equivalent of Proposition 1 in Martimort & Sand-Zantman (2016), but for the case of a pure global pollutant and two countries.

Proof See Online Appendix A2. □

The first-best agreement is feasible only if the level of uncertainty (γ) is relatively low. To understand this result, we need to figure out the impact of the length of the support R_θ on individual rationality and incentive compatibility. By Lemma 1 and equation (3), we know that when γ is high, incentives problems are more severe and truth-telling from low-cost countries may require very large information rent. As the budget balance constraint must always be satisfied, the compensations granted to these low-cost countries are limited by the necessity to ensure participation of all countries, including the ones with very high abatement costs. Therefore, when the support of the distribution of types is very large, one cannot find incentive compatible transfers that implement the first-best abatement levels and that give all types (including those with very high abatement costs) strictly more than their expected non-cooperative payoffs.

Condition (7) also implies that it is easier to implement the first-best agreement when θ_M is high. For a given γ, a larger θ_M means that, on average, abatement costs are lower and the welfare gains from reaching the first-best agreement are higher. These larger gains can be redistributed among countries to ensure incentive compatibility and individual rationality.

3 The two-stage game

As shown in Proposition 1, there is a tension between incentive compatibility, budget balance and individual rationality that may prevent countries from reaching the first-best agreement. In this section, I examine whether a certification stage preceding negotiations about abatement levels and transfers constitutes a channel to alleviate this tension and restore the feasibility of the first-best agreement. Specifically, I analyze a game consisting of two stages:

- *Stage 1 or certification stage.* Countries don't know the exact cost of abatement. They only have common prior beliefs over the support interval R_θ. To investigate what could be the cost of reducing pollution, countries have two options: doing the research privately and independently or sharing gathered information through an international agency, which must be able to monitor and certify the information transmitted (assisted by a scientific body for example). Let s_i denote the action (or strategy) of country i in stage 1. Ex-ante (i.e. before learning their own abatement costs), countries decide simultaneously whether to reveal their type through the certification device offered by the international agency ($s_i = C$) or to stay privately informed ($s_i = NC$).[12]

- *Stage 2 or abatement stage.* Countries at least privately know their types θ_i and they negotiate an environmental agreement $y^k = (a_1^k, a_2^k, t_1^k, t_2^k)$.

The game is solved backward for the collection of support intervals $R_\theta \times R_\theta$ characterized by:

$$\Theta = \left\{ R_\theta \times R_\theta \,\middle|\, R_\theta = [\theta_M - \gamma, \theta_M + \gamma] \text{ and } \frac{\theta_M}{3} < \gamma < \theta_M \right\}$$

i.e. support intervals R_θ such that the first-best agreement is not feasible without certification (condition (7) in Propostion 1 is not satisfied).

Stage 2: abatement game

The information structure (i.e. whether countries' types are privately or publicly known) will determine the kind of environmental agreement that can be implemented

[12]The assumption that the decision is taken ex-ante is made to abstract from signaling issues (as in Kakeu & Johnson (2018)). This specification is also consistent with the large literature on information exchange in oligopoly with private information about costs. In a typical scenario, the firms participate in information exchange before playing a one-shot Cournot game. Information is assumed to be verifiable, i.e. a firm can conceal its private information but cannot misrepresent it. Examples include Li (1985), Gal-Or (1986), and, more recently, Amir *et al.* (2010). I discuss this assumption in the next section.

in stage 2. In particular, three information structures are possible given the actions taken in stage 1.

1. Complete information when actions taken in stage 1 are $(s_1, s_2) = (C, C)$
2. Two-sided asymmetric information when actions taken in stage 1 are $(s_1, s_2) = (NC, NC)$
3. One-sided asymmetric information when actions taken in stage 1 are either $(s_1, s_2) = (C, NC)$ or $(s_1, s_2) = (NC, C)$

Below, we derive the optimal environmental agreements in these three possible situations.

Complete information. When countries' types θ_i are public knowledge, abatement and transfer levels do not have to be incentive compatible. The optimal mechanism will be the mechanism that maximizes global welfare, subject to the budget balance and individual rationality constraints (4) and (6). Using the participation constraint (6) evaluated at the first-best abatement levels, a_i^{FB}, the transfers necessary to make each country willing to participate in the first-best agreement are given by: $t_i^{FB}(\theta_i, \theta_j) \geq \frac{1}{8}\theta_i - \frac{1}{4}\theta_j$ for $i \neq j$. Summing up these two participation constraints, and using the budget balance constraint, $t_1^{FB}(\theta_1, \theta_2) + t_2^{FB}(\theta_1, \theta_2) = 0$, yield the condition:

$$0 \geq -\frac{1}{8}(\theta_1 + \theta_2)$$

which holds for all $R_\theta \times R_\theta \in \Theta$.

Lemma 2 *Under complete information, for all $R_\theta \times R_\theta \in \Theta$, the optimal mechanism is the first-best agreement y^{FB}.*

Therefore, when the actions taken in stage 1 are $(s_1, s_2) = (C, C)$, such that there is complete information in stage 2, the first-best mechanism is implemented

and country i's payoff is

$$\frac{1}{2}\theta_j + t_i^{FB}(\theta_i, \theta_j) \tag{8}$$

where $t_i^{FB}(\theta_i, \theta_j)$ satisfies the participation and budget balance constraints for $i \neq j$.

Two-sided asymmetric information. When countries' types are private information with support intervals $R_\theta \times R_\theta \in \Theta$, the optimal mechanism is the mechanism that maximizes the expected global welfare subject to (2), (4), and (6). Due to the tension between ensuring incentive compatibility for the low-cost countries and participation from the high-cost countries, the optimal mechanism cannot be the first-best agreement. To solve this tension, the optimal second-best mechanism reduces abatements below the first-best levels for all types (except for the lowest-cost country, $\theta_M + \gamma$). As shown in the Online Appendix A3, it is very difficult to find closed-form expressions for the second-best transfers and abatement levels. For this reason, I follow Martimort & Sand-Zantman (2016) who show that due to the convexity of the optimal second-best mechanism, this schedule can be approximated by a menu, $y^{SB} = (a_1^{SB}, a_2^{SB}, t_1^{SB}, t_2^{SB})$, with two options, denoted by SB1 and SB2.[13]

In option 1 (SB1), a country does not expand abatement effort beyond the non-cooperative level, i.e. $a_i^{SB1} = a_i^N = \frac{1}{2}\theta_i$ but it contributes a fixed amount $\underline{t}$ to a fund. In option 2 (SB2), a country chooses its first-best level of abatement $a_i^{SB2} = a_i^{FB} = \theta_i$ and receives a transfer made of two parts: a fixed contribution to a fund $\bar{t}$ (with $\bar{t} > \underline{t}$) and a subsidy of $\frac{1}{2}$ per unit of abatement. Similarly to the first-best transfer, this subsidy can be seen as a per-unit Pigouvian subsidy because $\frac{1}{2}$ is just the difference between the social and private marginal benefits of abatement.

Countries will self-select between these two options according to their costs of abatement. In the Online Appendix A3, I show that there exists a unique cutoff type θ^*, which is just indifferent between the two options, i.e. for which the expected

[13] Using numerical simulations, Martimort & Sand-Zantman (2016) also show that when the distribution of types is a uniform, the welfare loss from using a two-item menu instead of the optimal mechanism is very small.

payoff under each option is the same:

$$\underbrace{\frac{1}{8}\theta^* + \frac{1}{2}E_{\theta_j}[a_j^{SB}] - \underline{t}}_{Exp.\ Utility\ under\ SB1} = \underbrace{\frac{1}{2}\theta^* + \frac{1}{2}E_{\theta_j}[a_j^{SB}] - \bar{t}}_{Exp.\ Utility\ under\ SB2} \Leftrightarrow \theta^* = \frac{8}{3}(\bar{t} - \underline{t}) \tag{9}$$

Where $E_{\theta_j}[a_j^{SB}]$ is the expected abatement level of the other country under the two-item menu. All types below (resp. above) θ^*, will choose the first (resp. second) option. To ensure participation of the high-cost countries ($\theta_i < \theta^*$), their contribution $\underline{t}$ must be chosen such that they do not get an expected payoff lower than under the non-cooperative equilibrium (see equation (5)):

$$\frac{1}{8}\theta_i + \frac{1}{2}E_{\theta_j}[a_j^{SB}] - \underline{t} = U_i^N(\theta_i) \Leftrightarrow \underline{t} = \frac{1}{2}E_{\theta_j}[a_j^{SB} - a_j^N] \tag{10}$$

Intuitively, the contribution of high-cost countries must be equivalent to the externality gain created by the extra abatement effort under the second-best mechanism compared to the non-cooperative equilibrium. The system of equations formed by the budget balance constraint (4), the participation constraint (10) and the indifference condition (9) has a unique solution:

Lemma 3 *Under two-sided asymmetric information, for all $R_\theta \times R_\theta \in \Theta$, the optimal second-best mechanism $y^{SB} = (a_1^{SB}, a_2^{SB}, t_1^{SB}, t_2^{SB})$ is approximated by a two-item menu (SB1,SB2), with a cutoff type $\theta^* = \frac{\theta_M + \gamma}{2}$ such that:*

$$a_i^{SB} = \begin{cases} \frac{1}{2}\theta_i & if \quad \theta_i \in [\theta_M - \gamma, \theta^*) \\ \theta_i & if \quad \theta_i \in [\theta^*, \theta_M + \gamma] \end{cases}$$

And

$$t_i^{SB} = \begin{cases} -\underline{t} = -\frac{3(\theta_M+\gamma)^2}{64\gamma} & if \quad \theta_i \in [\theta_M - \gamma, \theta^*) \\ \frac{1}{2}\theta_i - \bar{t} = \frac{1}{2}\theta_i - \frac{3(\theta_M+\gamma)^2}{64\gamma} - \frac{3(\theta_M+\gamma)}{16} & if \quad \theta_i \in [\theta^*, \theta_M + \gamma] \end{cases}$$

Proof See Online Appendix A3. □

Therefore, when the actions taken in stage 1 are $(s_1, s_2) = (NC, NC)$, such that each country is privately informed about its own type in stage 2, country i's expected payoff from the second-best mechanism is:

$$\begin{cases} U_i^N(\theta_i) & \text{if} \quad \theta_i \in [\theta_M - \gamma, \theta^*) \\ U_i^N(\theta_i) + \frac{3}{8}(\theta_i - \theta^*) & \text{if} \quad \theta_i \in [\theta^*, \theta_M + \gamma] \end{cases} \tag{11}$$

A country with a type $\theta_i < \theta^*$ chooses the first option of the two-item menu and is left with its expected non-cooperative payoff. As its abatement costs are too high, this country would rather contribute to a fund that can be used to subsidize lower-cost countries (i.e. $\theta_i \geq \theta^*$). In return, this country does not have to abate beyond its non-cooperative level. Under the second option, the low-cost countries receive an information rent, which is increasing in their own type, and gives them an incentive to abate at their first-best level. This allows them to reach an expected utility higher than the non-cooperative level.

One-sided asymmetric information. Assume that the actions taken in stage 1 are $(s_1, s_2) = (C, NC)$ (the case where $(s_1, s_2) = (NC, C)$ is symmetric). Country 1's type is public knowledge and country 2's type is privately known. I first derive the following Lemma, which states that, under one-sided asymmetric information, the first-best agreement can be implemented if country 1's abatement cost is sufficiently low.

Lemma 4 *When country 1's type is public knowledge and country 2's type is privately known, the first-best agreement is implementable for all $\theta_2 \in R_\theta$ if and only if $\theta_1 \geq \tilde{\theta} = \max\{3\gamma - \theta_M; \theta_M - \gamma\}$.*

Proof See Online Appendix A4. □

The intuition behind this result is similar to Proposition 1. As θ_1 increases, abatement costs of country 1 decrease and the welfare gain from reaching the first-best agreement increases. Moreover, as country 1's type is public knowledge, the transfer to country 1 does not have to be incentive compatible. It is therefore easier to find transfers satisfying the participation constraints for both countries and that are incentive compatible for country 2 only.

The critical type $\tilde{\theta}$, from which the first-best agreement can be implemented, is increasing in the length of the support interval: the tension between incentive compatibility, budget-balance and individual rationality is stronger for higher levels of uncertainty. A higher value of θ_1 is then necessary to solve the tension. If the level of uncertainty is not too high (i.e. $\frac{\theta_M}{3} < \gamma \leq \frac{\theta_M}{2}$), then any country 1's type will allow to implement the first-best abatement levels under one-sided asymmetric information ($\tilde{\theta} = \theta_M - \gamma$). When $\frac{\theta_M}{2} < \gamma < \theta_M$, the critical type $\tilde{\theta} = 3\gamma - \theta_M$ is not the lower bound of the support interval.

If country 1's abatement cost is not sufficiently low ($\theta_1 < \tilde{\theta}$), countries have to negotiate a second-best agreement that maximizes the expected global welfare subject to the incentive, participation and budget balance constraints. The objective is to find a feasible agreement in which abatement levels are as close as possible to the first-best levels. As its abatement cost is publicly known, country 1 is requested to exert its first-best abatement effort and receives a transfer to ensure its participation, while country 2 is offered a two-item menu (SB1, SB2) similar to the one derived under two-sided asymmetric information.[14]

Lemma 5 *Under one-sided asymmetric information (with type θ_1 publicly known), for all $R_\theta \times R_\theta \in \Theta$, the optimal mechanism $y^*(\theta_1, \theta_2)$ is:*

- *the first-best agreement if $\theta_1 \geq \tilde{\theta}$: $y^* = y^{FB} = (a_1^{FB}, a_2^{FB}, t_1^{FB}, t_2^{FB})$*

[14]Note that here the optimal mechanism is the mechanism that maximizes the total expected welfare. This is different from Helm & Wirl (2016a), who look at a situation with one-sided private information in which the non-informed country has all the bargaining power and proposes a contract to the informed country.

where $a_i^{FB} = \theta_i$*, for* $i = 1, 2$ *and the budget-balanced transfers are:*[15]

$(t_1^{FB}, t_2^{FB}) = \left(\frac{\theta_1}{4} - \frac{1}{8}(3\gamma + \theta_M), \frac{1}{2}[\theta_2 - (\theta_M - \gamma)] + \frac{\theta_M - \gamma}{8} - \frac{\theta_1}{4}\right)$

- *the second-best agreement if* $\theta_1 < \tilde{\theta}$*:* $y^* = y^{SB} = (a_1^{SB}, a_2^{SB}, t_1^{SB}, t_2^{SB})$

  where country 1 exerts an abatement effort equal to its first-best level $a_1^{SB} = \theta_1$ *and receives an expected transfer* $t_1^{SB} = \frac{1}{4}\theta_1 - \frac{1}{8}(\theta_M - \gamma) - \frac{1}{16\gamma}((\theta_M + \gamma)^2 - \theta^{*2})$*, while country 2 is offered a two-item menu (SB1,SB2), with cutoff type* $\theta^* \in [\theta_M - \gamma, \frac{\theta_M + \gamma}{2}]$ *given by:*

$$\theta^* = \frac{3}{4}(\theta_M + \gamma) - \sqrt{\frac{(\theta_M + \gamma)^2}{16} + \gamma(\theta_M - \gamma)} \tag{12}$$

Proof See Online Appendix A4. □

The main difference compared to the optimal mechanism under two-sided asymmetric information is that the cutoff θ^* is lower. As country 1's type is public knowledge, incentives problems are less severe and it is possible to design transfers such that the second option of the menu, in which country 2 abates at the first-best level, is chosen by a larger proportion of country 2's types.

Using Lemma 5, country 1's expected payoff is:

$$\begin{cases} U_1^N(\theta_1) + \frac{1}{8}(\theta_1 - \theta_M + \gamma) & \text{if} \quad \theta_1 \in [\theta_M - \gamma, \tilde{\theta}) \\ U_1^N(\theta_1) + \frac{1}{8}(\theta_1 - 3\gamma + \theta_M) & \text{if} \quad \theta_1 \in [\tilde{\theta}, \theta_M + \gamma] \end{cases} \tag{13}$$

[15] As shown by expression (3), an incentive compatible first-best transfer will depend on the transfer received by the lowest type, $t_2^{FB}(\theta_1, \theta_M - \gamma)$. The objective of the next section is to determine whether countries have an incentive to undergo certification. The answer to that question will obviously depend on the transfer received under the first-best agreement. Here, we will consider the most favorable case for country 1 (i.e. the country choosing certification): country 2 (with private information) receives the minimum transfer required to ensure truthful participation, i.e. a transfer similar to expression (3) in which the participation constraint binds for the lowest type. As a result, the maximum transfer country 1 can expect under one-sided asymmetric information is t_1^{FB}, which satisfies its participation constraint as long as $\theta_1 \geq \tilde{\theta}$.

while country 2's payoff is

$$\begin{cases} U_2^N(\theta_1,\theta_2) & \text{if} \quad \theta_1 \in [\theta_M-\gamma,\tilde{\theta}),\ \theta_2 \in [\theta_M-\gamma,\theta^*) \\ U_2^N(\theta_1,\theta_2)+\frac{3}{8}(\theta_2-\theta^*) & \text{if} \quad \theta_1 \in [\theta_M-\gamma,\tilde{\theta}),\ \theta_2 \in [\theta^*,\theta_M+\gamma] \\ U_2^N(\theta_1,\theta_2)+\frac{3}{8}(\theta_2-\theta_M+\gamma) & \text{if} \quad \theta_1 \in [\tilde{\theta},\theta_M+\gamma] \end{cases} \tag{14}$$

where $\tilde{\theta}=\max\{3\gamma-\theta_M;\theta_M-\gamma\}$, θ^* is given by (12) and $U_2^N(\theta_1,\theta_2)\equiv V_2(y^N(\theta_1,\theta_2)|\theta_1,\theta_2)$ accounts for the fact that country 2 knows country 1's type. Even if country 1 is rewarded less per unit of effort, it gains from joining the agreement for all $\theta_1>\theta_M-\gamma$ (see equation (13)). By contrast, the privately informed country will obtain a payoff larger than its non-cooperative level only if $\theta_1\geq\tilde{\theta}$ or if it chooses option 2 of the second-best mechanism (when $\theta_1<\tilde{\theta}$).

Stage 1: certification

At this stage, countries simultaneously choose whether they agree to gather information through a joint research project and allow the international agency to certify their type: $s_i\in\{C,NC\}$. The effect of certification is twofold. On the one hand, by revealing a country's type, certification reduces the information asymmetry, which was responsible for the tension between incentive compatibility, budget balance and individual rationality constraints in stage 2. As shown in Lemmas 2 and 5, certification by country i implies that this country always exerts the first-best level of effort and country j chooses the first-best abatement level "more often". Certification generates efficiency gains defined as the difference in terms of global welfare between the first-best agreement and the optimal mechanism implemented in stage 2.

On the other hand, certification implies the loss of the information rent for the country revealing its type. Countries might have an incentive to free-ride on each others' certification in order to keep their potentially large information rent in the

second stage of the game. The level of uncertainty about abatement costs, γ, will play a key role in the decision of country i to share private information because it will affect the magnitude of the two effects mentioned above (through its impact on the cutoff types $\tilde{\theta}$ and θ^*, defined in Lemmas 3, 4 and 5).

First, the efficiency gains generated by either one-sided or two-sided certification are increasing in γ. Indeed, when γ is relatively small (close to $\frac{\theta_M}{3}$), both cutoff types $\tilde{\theta}$ and θ^* are either equal or close to the lower bound of the support interval, $\theta_M - \gamma$, which implies that most types will be required to abate at their first-best level under the optimal mechanism. The expected global welfare will be very similar under the three possible information structures. By contrast, as γ increases, both $\tilde{\theta}$ and θ^* increase, reducing the probability that a country staying privately informed chooses the first-best level of abatement under the optimal mechanism. Certification by this country will therefore generate higher efficiency gains.

Second, the probability to be considered as a low-cost country entitled to an information rent, and to obtain an expected payoff higher than under the outside option, $[1 - F(\theta^*)]$, is decreasing in the level of uncertainty. This effect reduces the incentives to free-ride on each others' certification. Payoffs associated with each second-stage outcome are given by equations (8), (11), (13) and (14). To find the equilibrium in stage 1, I first analyze countries' best responses.

Best response of country i ***when*** $s_j = NC$

In Lemma 6, I show that for relatively low levels of uncertainty, the second effect (loss of information rent) dominates the first one (efficiency gains). The efficiency gains generated by unilateral certification are not large enough. Even when the transfer granted to country j is the minimum transfer (i.e. that just satisfies individual rationality for the lowest type), the first-best transfer received by country i (when it chooses $s_i = C$) does not compensate for the loss of the information rent. Hence, $s_i = C$ cannot be a best response to $s_j = NC$. As efficiency gains are increasing in

γ, while the probability to obtain an information rent in stage 2 is decreasing in γ, there exists a level of uncertainty from which we can find transfers that compensate for the loss of a "less likely" information rent and $s_i = C$ is a best response to $s_j = NC$.

Lemma 6 *The best response of country i when $s_j = NC$ is the following:*

$$s_i^* = \begin{cases} NC & \text{if} \quad \gamma < \bar{\gamma}^{NC} = 0.9115\theta_M \\ C & \text{if} \quad \gamma \geq \bar{\gamma}^{NC} = 0.9115\theta_M \end{cases}$$

Proof See Online Appendix A5. □

Best response of country i when $s_j = C$

As mentioned earlier, under unilateral certification by country j, this country is already exerting the first-best level of effort, while country i's expected abatement level is also higher than under two-sided asymmetric information. As a result, efficiency gains generated by country i's certification will be smaller than in the situation where country j is privately informed (Lemma 6). When $s_j = C$, the threshold from which the efficiency gains will be large enough to find first-best budget-balanced transfers that compensate country i for the loss of the information rent it could obtain by staying privately informed, will therefore be higher than when $s_j = NC$.

Lemma 7 *The best response of country i when $s_j = C$ is the following:*

$$s_i^* = \begin{cases} NC & \text{if} \quad \gamma < \bar{\gamma}^{C} = 0.9296\theta_M \\ C & \text{if} \quad \gamma \geq \bar{\gamma}^{C} = 0.9296\theta_M \end{cases}$$

Proof See Online Appendix A6. □

The equilibrium in stage 1

It is now possible to characterize the equilibrium of the game in stage 1. Given

the best responses detailed above (Lemma 6 and Lemma 7), we have the following proposition:

Proposition 2 *The equilibrium in stage 1 is defined as follows:*

- *For* $\gamma \in \left(\frac{1}{3}\theta_M, \bar{\gamma}^{NC}\right)$*, where* $\bar{\gamma}^{NC} = 0.9115\theta_M$*, there is a unique pure-strategy Nash equilibrium* $(s_1^*, s_2^*) = (NC, NC)$*.*

- *For* $\gamma \in \left[\bar{\gamma}^{NC}, \bar{\gamma}^{C}\right)$*, where* $\bar{\gamma}^{C} = 0.9296\theta_M$*, there are two pure-strategy Nash equilibria* $(s_1^*, s_2^*) = (C, NC)$ *and* $(s_1^*, s_2^*) = (NC, C)$ *and one mixed-strategy equilibrium, with the probability that country* i *chooses* $s_i^* = C$ *given by:*

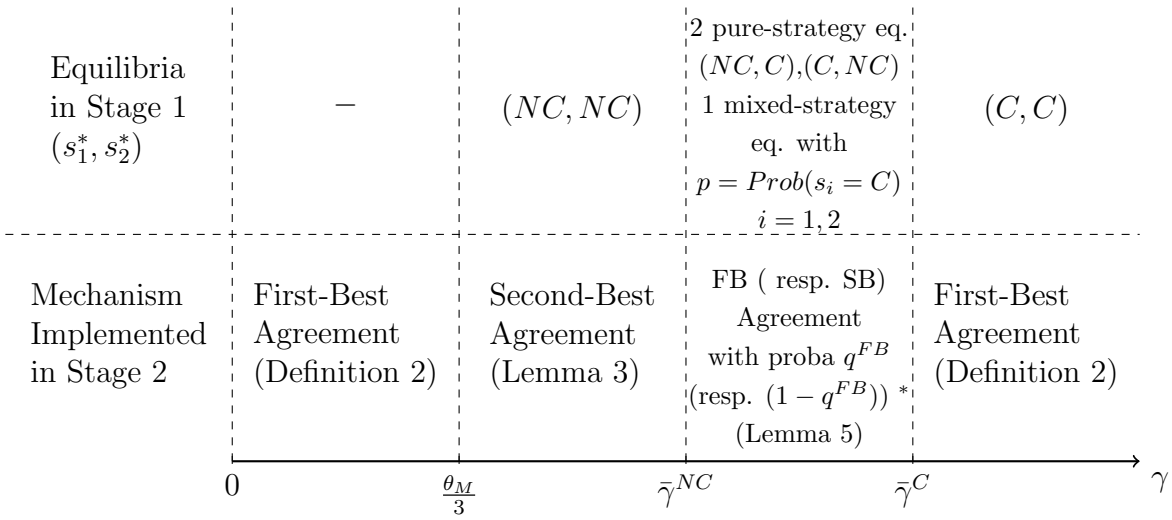
$$p = \frac{1}{1 + \frac{16\gamma^2(2\theta_M - 3\gamma) + 12(2\gamma - \theta_M)(\theta_M + \gamma - \theta^*)^2}{16\gamma^3 - 32\gamma(\theta_M - \gamma)(2\gamma - \theta_M) - 3\gamma(\theta_M + \gamma)^2}} \in (0, 1)$$

- *For* $\gamma \in \left[\bar{\gamma}^{C}, \theta_M\right)$*, there is a unique pure-strategy Nash equilibrium* $(s_1^*, s_2^*) = (C, C)$*.*

Proof See Online Appendix A7. □

Can certification restore the feasibility of the first-best agreement?

The result of the two-stage game is summarized in Figure 1. The opportunity for countries to set up an international agency entitled to certify their abatement costs can fully restore the feasibility of the first-best agreement only if the level of uncertainty is very high, i.e. $\gamma \geq \bar{\gamma}^{C}$. For these levels of uncertainty, the second-best agreement is relatively close to the non-cooperative equilibrium in terms of welfare because the probability that a country chooses its non-cooperative level of abatement, $F(\theta^*)$, is high. The efficiency gains of reaching the first-best agreement will then be sufficiently large to design first-best budget-balanced transfers $t_i^{FB}(\theta_i, \theta_j)$ under complete information such that both countries have an incentive to undergo



* q^{FB} is the probability that the first-best is implemented when $\gamma \in (\bar{\gamma}^{NC}, \bar{\gamma}^{C})$.
(a) For the pure-startegy equilibria, the FB agreement is implemented if for $s_i^* = C$, $\theta_i > \tilde{\theta}$. This implies that $q^{FB} = 1 - F(\tilde{\theta}) = \frac{\theta_M - \gamma}{\gamma}$.
(b)For the mixed-startegy equilibrium, the FB agreement is implemented with probability $q^{FB} = p^2 + 2p(1-p)\frac{\theta_M - \gamma}{\gamma}$, where p is the probability that country i plays $s_i = C$.

Figure 1: Summary of the results

certification. No country has an incentive to free-ride on the other's action because the risk to implement a second-best agreement agreement, which is close to the non-cooperative equilibrium, is substantial and so a country's expected benefits of keeping its information rent are very low.

4 Discussion

For tractability, the model presented in sections 1-3 relies on a number of simplifying assumptions. In this section, I discuss the impact of relaxing some of them.

4.1 Asymmetric information about benefits from abatement

So far we have considered a model in which private information is about costs, while benefits from abatement are known. In section B of the Online Appendix, I shift private information from costs to benefits and show that the main results of the

model remain qualitatively unchanged.

It is first possible to derive the equivalent of Proposition 1. In other words, we can show that a first-best agreement will be implementable only for levels of uncertainty about marginal benefits that are relatively low. This is again due to the tension between ensuring incentive compatibility for countries with high marginal benefits (which have an incentive to misreport their type to reduce the effort required under the agreement) and individual rationality for low-type countries.[16]

As for the case of private information about costs, it is very difficult to find a closed-form solution for the second-best transfers and abatement levels.[17] At the same time, resorting to an approximation by a two-item menu in which countries choosing their non-cooperative abatement level, contribute a fixed (type-independent) amount $\underline{t}$ to a fund, as described in section 3, will not be appropriate. With private information about benefits, the optimal second-best mechanism is similar to the mechanism presented in Online Appendix A3: the participation constraint for countries with low marginal benefits from abatement binds and they exert their non-cooperative level of effort, while countries with higher marginal benefits strictly gain from joining the agreement and exert a higher level of effort. However, the expected transfer received by a country for which the participation constraint binds is not constant anymore. This is due to the fact that the externality gain generated by a second-best agreement (and that is used to compute the transfer under the second-best mechanism) depends on a country's marginal benefit, which is constant and equal to 1/2 when uncertainty is about costs, but is type-dependent and privately known when uncertainty is about marginal benefits.[18]

[16]Helm & Wirl (2016b) also discuss how a shift of private information from costs to benefits affects the conditions to implement the first-best agreements, but they focus on a case with a continuum of infinitesimally small countries.

[17]Helm & Wirl (2014) and Helm & Wirl (2016b) derive the optimal second-best mechanism when private information is about marginal benefits. However, they rely on a principal-agent framework with one-sided asymmetric information only.

[18]Martimort & Sand-Zantman (2016) show that in a case where countries also differ in terms of their marginal benefit of abatement and it is private information, the properties of their two-item menu are similar to those obtained when costs are the sole source of heterogeneity. However, they

Nevertheless, we can pursue a more modest objective and derive the equilibrium of a two-stage game in which if countries cannot implement the first-best agreement in stage 2, they resort to the non-cooperative equilibrium outcome. The optimal abatement efforts and transfers will obviously be affected, but the main result that both countries will share their information only if the level of uncertainty is very high, remains.

4.2 Verifiable Information

So far, we have assumed that in the certification stage, countries have to decide whether to conduct research on their abatement cost privately or through a joint research project and the information shared is costlessly and totally verifiable by an international agency. In sections C and D of the Online Appendix, I explore two extensions to show how these assumptions affect the main results of the model.

I first show in section C that introducing research costs (or costs of setting up an international agency in stage 1) only affects the threshold values obtained in section 3, $\bar{\gamma}^{NC}$ and $\bar{\gamma}^{C}$. However, as long as these research costs are not too high, the main result that there exist some support intervals, characterized by high levels of uncertainty, for which certification by both countries is the unique equilibrium in stage 1, still remains.

Second, instead of assuming that certification reveals each player's type and thus eliminates all private information, I suppose that the international agency can only certify whether the country's cost parameter is above or below the mean θ_M. Partial verifiability reduces the benefits, and so the attractiveness, of certification because it does not completely eliminate asymmetric information and does not restore the feasibility of the first-best for all possible support intervals and all types of countries. Partial verifiability also reduces the cost of certification as a country

derive this result assuming that the externality is common knowledge and constant, and only the local benefits are private information. This implies that the contribution under the first option of the menu is still constant.

using certification does not lose its entire information rent. This may increase a country's willingness to use this mechanism.

In section D of the Online Appendix, I show that the second effect dominates and the threshold values for the equilibrium in stage 1, $\bar{\gamma}^{NC}$ and $\bar{\gamma}^{C}$, are lower than in the model with full revelation of countries' types. Countries will use certification more often; however, certification by both countries does not imply that the first-best agreement will be implemented in stage 2 as it does not completely eliminate asymmetric information. In fact, when certification by both countries restore the feasibility of the first-best agreement for all possible types in $[\theta_M - \gamma, \theta_M + \gamma]$ (i.e. when $\gamma \leq \frac{1}{2}\theta_M$), the unique equilibrium is (NC, NC). By contrast, for $\gamma \in (\bar{\gamma}^{C}, \theta_M)$, where the unique equilibrium in stage 1 is (C, C), certification restores the feasibility of the first-best agreement only when both countries report low abatement costs in stage 1, i.e. $\theta_i \in [\theta_M, \theta_M + \gamma]$. In other cases, certification still generates efficiency gains as the reduction in information asymmetry allows countries to negotiate a second-best agreement in which a larger proportion of types chooses their first-best abatement levels.

4.3 Timing in stage 1

The timing of stage 1 (i.e. the fact that countries decide whether to share information ex ante and the transmitted information is certifiable or provable by the international institution) is used to abstract from signaling incentives. A more realistic setting would be that in the certification stage, a country is better informed about its own abatement costs than about the abatement costs of the other country. This would lead to an analysis of strategic information transmission, in which not accepting certification sends a signal to the other country.

The game discussed in this paper is however more complex than the games analyzed in the literature about strategic information transmission (e.g. Okuno-

Fujiwara *et al.* 1990; Cramton & Palfrey 1995, Hagenbach *et al.* 2014). These papers typically analyze pre-play communication in Bayesian games with a unique equilibrium. Even though the equilibrium abatement levels in stage 2 are unique, the first-best transfers are not necessarily unique. Moreover, the nature of the agreement negotiated in stage 2 (first-best or second-best) depends on the level of uncertainty, which will in turn depend on the updated beliefs about the other country's abatement costs after the communication stage. The resolution of the game then requires making strong assumptions regarding the beliefs about the outcome of future negotiations (nature of agreement and transfers levels) conditional on observable moves.

To overcome these issues and explore how changing the timing of information sharing affects the results of the paper, I present, in section E of the Online Appendix, a simplified version of the model, in which I use the concept of *credible veto set* (Cramton & Palfrey 1995) to determine the off-equilibrium path beliefs. In this new version of the model, countries privately know their type in stage 1 and have to decide to share this information through the certification device proposed by the international agency. When deciding to refuse the certification device in stage 1, a country will consider how the other country's beliefs (and so the agreement negotiated in stage 2) may change as a result of this decision.

I focus on the existence of a *ratifiable* certification device, i.e. whether for both countries at the same time and all types $\theta_i \in [\theta_M - \gamma, \theta_M + \gamma]$, there does not exist a subset of types that strictly benefit from refusing certification under updated beliefs. The full analysis is detailed in section E of the Online Appendix. The main result is that a certification device will be ratifiable only when the level of uncertainty is sufficiently high (i.e. $\gamma > \frac{2}{5}\theta_M$). Even though this looks similar to the results in section 3, the interpretation is very different because the change in timing affects the costs of refusing certification. In the main model, the cost of staying privately informed was a loss of efficiency (e.g. implementation of a second-best agreement instead of the first-best agreement). In this modified model, both

strategies (NC and C) imply the (partial) loss of a country's information rent due to updated beliefs.

As shown previously, the cost of certification might be substantial for low-cost countries (i.e. countries with a high θ_i) because they lose a potentially high information rent when they reveal their type. This is not the case for high-cost countries, because their payoffs under a second-best agreement will always be either equal or very close to the non-cooperative outcome. By refusing certification, country i sends a credible signal that it has a relatively low-cost of abatement $\theta_i \in [\underline{\theta}_i, \theta_M + \gamma]$, with $\underline{\theta}_i > \theta_M - \gamma$. These updated beliefs allow the mechanism designer to update the agreement in stage 2 and reduce the information rent for country i. But these updated beliefs also imply a lower level of asymmetric information.

For low levels of γ, even a small reduction in asymmetric information induced by updated beliefs would be enough to implement an agreement where both countries choose their first-best abatement levels. In that case, both ratifying and refusing certification lead to the same efficiency gains. It is then possible to find a subset of low-cost countries that are better off staying privately informed with updated beliefs and a (reduced) information rent than revealing their type through certification and losing their entire information rent. For higher levels of uncertainty, the incentives problems are more severe. As a result, it is impossible to find updated beliefs $[\underline{\theta}_i, \theta_M + \gamma]$ that generate a sizable increase in global welfare, while still allowing country i to benefit from a large information rent. In this situation, a country will ratify the certification device for all types.

5 Conclusion

This paper takes a mechanism design approach to study the effect of asymmetric information about abatement costs on the feasibility of an efficient environmental agreement when participation is voluntary. Due to the tension between incentive

compatibility and participation, a first-best agreement cannot always be reached. For this reason, I investigate the role of a pre-negotiation phase as an information-sharing and certification device to alleviate the inefficiencies generated by information asymmetry and voluntary participation. The introduction of this certification stage is motivated by the fact that many existing international environmental agreements with explicit abatement targets have been preceded by an umbrella convention that set up a structure to gather countries' information related to the environmental issue.

The ability of a certification stage to restore the feasibility of the first-best agreement will depend on the relative magnitude of two effects. First, by eliminating totally or partially information asymmetry between countries at the negotiation stage, certification reduces the tension between the incentive compatibility and participation constraints. Second, the country using certification loses the possibility to misreport its abatement cost during the negotiation stage. This loss is larger for countries with lower abatement costs, which under asymmetric information, will require a higher payoff to satisfy the incentive compatibility constraint. When the level of uncertainty about abatement costs is high enough, the welfare gains of reducing information asymmetries and reaching the first-best agreement are substantial, while the probability to be considered as a low-cost country and benefit from a large information rent is low. As a result, the first effect dominates and both countries opt for certification in the first stage. The feasibility of the first-best agreement is restored for those high levels of uncertainty.

The results of the model point towards the importance of coordinated research efforts to achieve efficient environmental agreements and prevent countries from claiming substantial information rents. However, information-sharing is not sufficient, this information must also be certified such that it can be used as a basis to negotiate subsequent environmental targets. For example, in the case of acid rains, the creation of mutually agreed-upon scientific knowledge has been a the center of

international cooperation since the very beginning (e.g. in the LRTAP Convention). This principle also appears in the second Sulfur Protocol in 1994. The national emission reductions under this protocol have been calculated using integrated assessment models (especially RAINS). These models have been developed through international research projects involving a large number of scientists from different countries. The results of these models were therefore less easy to dispute, reducing the ability of countries to misreport their willingness-to-pay for an agreement.

A lot of simplifying assumptions have been used to highlight the effects of certification. A first important extension could be to consider that abatement efforts are not totally observable, so that there is a problem of moral hazard during the implementation of the first-best agreement. Second, interim individual rationality is used as a participation constraint. In the case of international environmental treaties, countries can withdraw from the agreement at no cost after observing the outcome of the negotiations. This will clearly affect the ability of countries to design a mechanism implementing the first-best abatement levels and it would be worth to explore the impact of imposing ex-post, rather than interim, individual rationality constraints.

References

Amir, Rabah, Jin, Jim Y, & Troege, Michael. 2010. Robust results on the sharing of firm-specific information: Incentives and welfare effects. *Journal of Mathematical Economics*, **46**(5), 855–866.

Baliga, Sandeep, & Maskin, Eric S. 2003. Mechanism design for the environment. *Pages 305–324 of:* Maler, K.-G., & Vincent, J.R. (eds), *Handbook of Environmental Economics, Volume 1.* Amsterdam, North-Holland: Elsevier Science.

Benedick, Richard E. 1998. *Ozone Diplomacy: New Directions in Safeguarding the*

Planet. Cambridge, MA: Harvard University Press.

Börgers, Tilman, & Norman, Peter. 2009. A note on budget-balance under interim participation constraints: the case of independent types. *Economic Theory*, **39**(3), 477–489.

Brandt, Urs Steiner. 2004. Unilateral actions, the case of international environmental problems. *Resource and Energy Economics*, **26**, 373–391.

Brandt, Urs Steiner, & Nannerup, Niels. 2013. Unilateral actions as signals of high damage costs: distorting pre-negotiations emissions in international environmental problems. *Environmental Economics*, **4**(2), 31–41.

Caparros, Alejandro J., Pereau, Jean-Christophe, & Tazdait, Tarik. 2004. North-South Climate Change Negotiations: A Sequential Game with Asymmetric Information. *Public Choice*, **121**(3), 455–480.

Cramton, Peter C., & Palfrey, Thomas R. 1995. Ratifiable mechanisms: learning from disagreement. *Games and Economic Behavior*, **10**(2), 255–283.

Elofsson, Katarina. 2007. Cost Uncertainty and Unilateral Abatement. *Environmental and Resource Economics*, **36**, 143–162.

Espinola-Arredondo, Ana, & Munoz-Garcia, Felix. 2012. *Keeping Negotiations in the Dark: Environmental Agreements under Incomplete Information.* School of Economic Sciences Working Paper Series, 2010-20, Washington State University.

Finus, Michael, Ierland, Ekko van, & Dellink, Rob. 2006. Stability of Climate Coalitions in a Cartel Formation Game. *Economics of Governance*, **7**(3), 271–291.

Gal-Or, Esther. 1986. Information transmission—Cournot and Bertrand equilibria. *The Review of Economic Studies*, **53**(1), 85–92.

Hagenbach, Jeanne, Koessler, Frédéric, & Perez-Richet, Eduardo. 2014. Certifiable Pre-play Communication: Full Disclosure. *Econometrica*, **82**(3), 1093–1131.

Helm, Carsten, & Wirl, Franz. 2014. The principal-agent model with multilateral externalities: An application to climate agreements. *Journal of Environmental Economics and Management*, **67**(3), 141–154.

Helm, Carsten, & Wirl, Franz. 2016a. Climate policies with private information: The case for unilateral action. *Journal of the Association of Environmental and Resource Economists*, **3**(4), 893–916.

Helm, Carsten, & Wirl, Franz. 2016b. Multilateral externalities: Contracts with private information either about costs or benefits. *Economics Letters*, **141**(1), 27–31.

Kakeu, Johnson, & Johnson, Erik Paul. 2018. Information Exchange and Transnational Environmental Problems. *Environmental and Resource Economics*, **71**(2), 583 – 604.

Konrad, Kai A., & Thum, Marcel P. 2014. Climate Policy Negotiations with Incomplete Information. *Economica*, **81**, 244–256.

Laffont, Jean-Jacques, & Maskin, Eric S. 1979. A differential approach to expected utility maximizing mechanisms. *Pages 289–308 of:* Laffont, Jean-Jacques (ed), *Aggregation and Revelation of Preferences.* Amsterdam, North-Holland: Elsevier.

Li, Lode. 1985. Cournot oligopoly with information sharing. *The RAND Journal of Economics*, 521–536.

Mailath, George J., & Postlewaite, Andrew. 1990. Asymmetric Information Bargaining Problems with Many Agents. *Review of Economic Studies*, **57**, 351–367.

Martimort, David, & Sand-Zantman, Wilfried. 2016. A Mechanism Design Approach to Climate Agreements. *Journal of the European Economic Association*, **14**(3), 669–718.

McAfee, R. Preston, & Reny, Philip J. 1992. Correlated Information and Mecanism Design. *Econometrica*, **60**(2), 395–421.

Myerson, Roger B. 1982. Optimal coordination mechanisms in generalized principal-agent models. *Journal of Mathematical Economics*, **10**(1), 67–81.

Myerson, Roger B., & Satterthwaite, Mark A. 1983. Efficient Mechanisms for bilateral trading. *Journal of Economic Theory*, **29**(2), 265–281.

Neeman, Zvika. 2004. The relevance of private information in mechanism design. *Journal of Economic Theory*, **117**(1), 55–77.

Nordhaus, William D., & Yang, Zili. 1996. A Regional Dynamic General-Equilibrium Model of Alternative Climate-Change Strategies. *American Economic Review*, **86**(4), 741–765.

Okuno-Fujiwara, Masahiro, Postlewaite, Andrew, & Suzumura, Kotaro. 1990. Strategic information revelation. *The Review of Economic Studies*, **57**(1), 25–47.

Rob, Rafael. 1989. Pollution Claim Settlements under Private Information. *Journal of Economic Theory*, **47**(2), 307–333.

Wagner, Ulrich J. 2001. The design of stable international environmental agreements: economic theory and political economy. *Journal of Economic Surveys*, **15**(3), 377–411.